CONSISTAN: Jurnal Tadris Matematika
Volume 4, Number 01, 2026, 137-148
ISSN (Online): 3025-0943
https://ejournal.alqolam.ac.id/index.php/CONSISTAN
Diterima: 14-05-2026 *Direvisi: 22-05-2026* *Disetujui: 30-05-2026*

Item Analysis of an Adaptive Assessment Question Bank in Maritime Vocational Education

Tika Septia[1], Dirhamsyah[2], Tanti Diyah Rahmawati[3], Rezky Rahma Ruslan[4]
[1,2,3]*Teknologi Rekayasa Permesinan Kapal,* Politeknik Pelayaran Surabaya, Indonesia.
[4]*Teknologi Rekayasa Operasi Kapal*, Politeknik Pelayaran Surabaya, Indonesia.
e-mail: tika.septia@poltekpel-sby.ac.id,

ABSTRACT
Adaptive assessment systems depend on question banks whose difficulty rises in a credible way and whose items can distinguish learner performance. This study reports the design and validation of a 90-item question bank for adaptive learning in maritime mathematics at Politeknik Pelayaran Surabaya. The bank covered six STCW-aligned topics (speed-distance-time, ETA calculation, compass correction, great circle navigation, tidal height, altitude correction), arranged them across three progressive difficulty levels, and linked them to an AI-tutor system using a 2-up/2-down adaptive staircase algorithm. Validation had two parts. The maritime mathematics curriculum committee first conducted an expert pedagogical audit with a rubric covering curriculum alignment, difficulty progression, answer correctness, and linguistic clarity. Simulated pilot testing with synthetic learner profiles was then used to check whether the adaptive engine behaved as expected. In that audit, 15 of 90 items needed revision, mainly for clarity and difficulty calibration, while 75 were retained without modification. After revision, the bank still showed a balanced spread across topics and difficulty levels and maintained appropriate content validity. Simulation results likewise confirmed the intended pattern: difficulty shifted only after two consecutive correct or incorrect responses. For that reason, expert review paired with simulation offers a workable validation route for adaptive assessment systems in specialized vocational education contexts, especially when large-scale learner pilot testing is not yet feasible.

Keywords: *question bank validation; adaptive learning; maritime mathematics; item analysis; vocational education*

INTRODUCTION

Maritime polytechnics that prepare future merchant navy officers need assessment systems that stay aligned with international competency standards. Through the Standards of Training, Certification and Watchkeeping for Seafarers (STCW), the International Maritime Organization (IMO) sets those minimum requirements and explicitly includes mathematical and navigational calculations as essential competencies for deck and engineering officers (International Maritime Organization, 2017). Maritime mathematics is especially demanding

in that setting. It joins abstract mathematical concepts with applied navigational contexts such as great circle sailing, compass correction, tidal calculations, and celestial navigation.

At their core, adaptive learning systems handle differences in how learners prepare and perform. Rather than forcing every student down the same assessment path, these platforms adjust to individual needs in real time. This fluid adjustment sustains the balance between keeping the student focused and driving actual learning (VanLehn, 2011; Aleven et al., 2017). None of this works, however, without a solid question bank. The pool needs a clear progression in difficulty, questions that can distinguish strong performers from weaker ones, and tight alignment with curricular demands (Hambleton & Swaminathan, 1985).

Recent discussions in the literature point to similar priorities. Modern LLM development often relies on retrieval mechanisms alongside grounding and factuality checks to reduce hallucination risks and keep system outputs traceable. Beneath these technical safeguards, however, there remains a classic measurement problem: is the test question actually capturing the construct it was designed to evaluate? For an assessment tool operating in a specialized domain, relying on one metric alone is not sufficient. Researchers generally look for converging evidence and triangulate multiple measures, frequently cross-referencing content validity indices against inter-rater reliability coefficients or generalizability theory. Taken together, this literature supports the proposed setup as a domain-specific application rather than a standard open-ended chatbot.

While adaptive learning environments have become fairly common in general education, validated question pools for specialized vocational fields remain difficult to find. Maritime mathematics illustrates this gap clearly. Drafting items here is not just about numbers; authors must weave computations into realistic navigational contexts, use terminology accurately, and reflect scenarios cadets are likely to encounter in practice. Without validated item pools, adaptive learning in maritime vocational education remains difficult to deploy.

Several studies have addressed question bank validation in general educational contexts. Pane et al. (2017) demonstrated that personalized learning implementations with well-calibrated item pools produce measurable learning gains. Kulik and Fletcher (2016) meta-analyzed 50 intelligent tutoring systems and found effect sizes strongly correlated with item bank quality. For adaptive systems specifically, Levitt (1971) established the theoretical foundation for the 2-up/2-down staircase algorithm, which converges toward a learner's current competency threshold without requiring prior item calibration data. Hambleton and Swaminathan (1985) outlined standard procedures for item analysis, including difficulty index (p-value), discrimination index (D), and distractor analysis.

In the Indonesian context, research on adaptive assessment has primarily focused on general academic subjects in higher education, with limited attention to maritime vocational training. Wahyuwono et al. (2024) developed NLP-based assistants for academic services but did not address assessment item quality. Research gaps therefore remain in: (1) systematic validation methodology for small-scale specialized question banks, (2) integration of validated items with adaptive algorithms in maritime mathematics, and (3) practical validation approaches when large-scale learner pilots are not yet available.

This study addresses these gaps by presenting a systematic validation approach for a 90-item maritime mathematics question bank developed at Politeknik Pelayaran Surabaya. The validation combines an expert pedagogical audit with simulation-based verification of adaptive engine behavior. The research contributes: (1) a documented question bank structure covering six STCW-aligned topics across three difficulty levels, (2) a structured rubric for expert validation of maritime mathematics items, and (3) evidence of adaptive engine correctness through simulation testing. The approach is practical for vocational institutions developing adaptive systems in specialized domains where subject matter experts are available, but large learner populations for pilot testing are constrained.

METHOD

This research employed a Design and Development (DDR) methodology (Richey & Klein, 2007) focused on systematic validation of an assessment instrument. The research proceeded through four phases: (1) question bank specification based on STCW curriculum analysis, (2) item development by subject matter experts, (3) expert pedagogical audit with structured rubric, and (4) simulation-based verification of adaptive engine integration. Formal learner pilot testing was positioned as future work pending institutional ethical approval and cadet scheduling.

Question Bank Specification. The question bank architecture was designed to support an adaptive learning system employing a rule-based 2-up/2-down staircase algorithm (Levitt, 1971). The algorithm requires items organized by progressive difficulty levels, with sufficient items per level to prevent repetition during typical learner sessions. The specification defined six topics corresponding to core STCW navigational mathematics competencies, three difficulty levels reflecting cognitive complexity, and five items per topic-difficulty combination yielding 90 total items. Table 1 presents the complete specification.

Table 1. Question Bank Specification Matrix

No	Topic (STCW Reference)	Difficulty 1	Difficulty 2	Difficulty 3	Total
1	Speed, Distance, Time (Table A-II/1)	5	5	5	15
2	ETA Calculation (Table A-II/1)	5	5	5	15

3	Compass Correction (Table A-II/1)	5	5	5	15
4	Great Circle Navigation (Table A-II/2)	5	5	5	15
5	Tidal Height (Table A-II/1)	5	5	5	15
6	Altitude Correction (Table A-II/2)	5	5	5	15
	Total	30	30	30	90

We operationalized difficulty using Bloom's revised taxonomy (Anderson & Krathwohl, 2001) and Webb's Depth of Knowledge framework (Webb, 1997). Difficulty 1 items target the Apply level and require direct use of a single formula with given values. Difficulty 2 items move to the Analyze level, requiring multi-step reasoning with unit conversions or intermediate calculations. Difficulty 3 items target the Evaluate/Create level by presenting integrated scenarios that require synthesis across multiple concepts.

Item Development. Three instructors with an average of 8 years of maritime navigation teaching experience authored the items. Each item followed a standardized structure: a navigational scenario that framed the problem, specific numerical values and conditions, a clearly stated question, and, where applicable, four multiple-choice alternatives. Standard nautical conventions governed the mathematical formulas and symbols. The items were written in Indonesian to match instructional practice at Politeknik Pelayaran Surabaya, while English technical terms familiar in maritime work were retained (e.g., ETA, compass bearing, variation).

Expert Pedagogical Audit. For the validation phase, we relied on a structured rubric drawn from established item analysis frameworks (Hambleton & Swaminathan, 1985; Brookhart, 2010). Our experts evaluated four dimensions. They checked whether an item aligned with the target STCW competency, whether its cognitive demand matched the assigned difficulty, and whether the mathematical calculations were correct. Lastly, they examined linguistic clarity. Reviewers checked that the wording was not ambiguous, that maritime terms were used appropriately, and that the reading level was suitable for incoming first-year cadets. Scores ran from 0 to 2. Within this framework, scoring a 2 meant full compliance. A 1 pointed to minor necessary revisions, while a 0 flagged the item for major revision or rejection. Any item that did not secure a 2 on even one dimension was pulled for closer review. Two senior instructors from the Nautical Science department reviewed the items independently and resolved disagreements through discussion. Inter-rater agreement was then estimated with Cohen's kappa.

Adaptive Engine Simulation. To test the adaptive engine, we programmed three synthetic learner profiles: novice, intermediate, and advanced. We fixed their respective probability of answering correctly at 0.30, 0.60, and 0.85. Under the 2-up/2-down rule, every profile completed 100 simulated sessions. Each session contained 10 items. The system logged

the starting level, the answer sequence, the points where difficulty increased or decreased, and the final tier the learner reached. We established clear expectations before running the simulation: novices should settle at level 1, intermediates should oscillate around level 2, and advanced cadets should move toward level 3.

System Implementation Context. On the technical side, this question bank powers an AI-tutor framework. The architecture pairs a FastAPI (Python 3.12) backend and SQLite database with a React 18 frontend, relying on KaTeX to handle the math rendering. If a cadet asked for a hint, the Gemini 2.5 Flash Large Language Model generated the step-by-step breakdown. Cadets accessed the deployed system through standard web browsers on a Virtual Private Server. Fuller implementation details of the LLM tutor component and overall system architecture are reported separately because this article concentrates on question bank validation**.**

RESULTS AND DISCUSSION

Expert Audit Results

The expert pedagogical audit yielded comprehensive scores across all four dimensions for the entire 90-item pool. The two senior reviewers showed solid alignment. Their calculated Cohen's kappa reached 0.78. This number lands comfortably in what is generally considered the range of substantial consensus (Landis & Koch, 1977). Table 2 provides a complete breakdown of these topic-level findings.

Table 2. Expert Audit Results by Topic

Topic	Items	Accepted	Minor Revision	Major Revision
Speed, Distance, Time	15	14	1	0
ETA Calculation	15	13	2	0
Compass Correction	15	12	3	0
Great Circle Navigation	15	11	3	1
Tidal Height	15	13	2	0
Altitude Correction	15	12	2	1
Total	90	75	13	2
Percentage	100%	83.3%	14.4%	2.2%

Of the 90 drafted questions, 75 items (83.3%) cleared the audit without any required changes. Another 13 items (14.4%) required minor adjustments. This left 2 items (2.2%) in need of major revision. The two major cases both appeared at Difficulty 3 level and involved integrated scenarios whose original reference answers contained computational errors. The reviewers identified these computational problems during the mathematics verification phase and then recalculated the scenarios to correct the reference answers.

Of the 13 cases needing minor revision, 9 involved phrasing. The reviewers encountered three recurring issues: unclear pronoun references, mixed metric and imperial units, and sentence structures that reduced readability. The other 4 flagged items involved difficulty calibration. Two were harder than their assigned level, while the other two were easier than expected. Following expert discussion, the items were either revised or moved to a more appropriate difficulty level.

Revised Question Bank Characteristics

After these corrections were implemented, the pool retained its structure. The final bank maintained all 90 items, with 5 questions in each of the 18 topic-difficulty intersections. To quantify content validity, we used the experts' final ratings to calculate Aiken's V coefficient (Aiken, 1985). That computation yielded a content validity index (CVI) of 0.87. Because this exceeds the standard 0.80 threshold (Polit & Beck, 2006), the bank's overall content validity is judged to be strong. As expected, curriculum alignment drew the highest average ratings (mean = 1.94 out of 2.0). This outcome reflects the use of STCW parameters during the drafting phase. Scores for difficulty calibration were more varied (mean = 1.78, SD = 0.34), which is reasonable because difficulty ratings remain partly subjective before actual student performance data are available.

Adaptive Engine Simulation Results

The numbers from the simulation were reassuring. They confirmed that the 2-up/2-down staircase mechanism behaved as designed for all three synthetic profiles. A summary of the 300 total simulated runs (100 per persona) is provided in Table 3.

Table 3. Adaptive Engine Simulation Results

Profile	p(correct)	Final Difficulty (Mean)	Final Difficulty (Mode)	Transitions/Session (Mean)
Novice	0.30	1.12	1	2.4
Intermediate	0.60	2.05	2	3.8
Advanced	0.85	2.78	3	2.1

The simulation results align with theoretical expectations. Novice profiles converged rapidly to difficulty 1 and remained there, demonstrating the algorithm's correct reduction behavior when two consecutive incorrect responses occur. Advanced profiles reached and stabilized at difficulty 3, confirming the escalation mechanism. Intermediate profiles showed the expected oscillation pattern around difficulty 2, with the highest number of transitions per session (3.8), consistent with the algorithm tracking a learner whose performance is near the decision boundary between levels.

An important behavioral observation was the transition latency. The 2-up/2-down algorithm requires two consecutive events to trigger a transition, creating a natural delay between performance changes and difficulty adjustment. In simulation, this produced an

average of 3.2 items between difficulty changes, which is considered pedagogically appropriate as it prevents oscillation due to single lucky or unlucky responses while maintaining responsiveness to genuine skill level shifts.

Discussion

The validation results demonstrate that the developed question bank meets established criteria for content validity and is suitable for integration with adaptive assessment algorithms. The 83.3% acceptance rate in expert audit is consistent with reported rates in similar vocational assessment development studies (Brookhart, 2010), supporting the quality of the initial item development process. The concentration of revisions in linguistic clarity rather than technical correctness suggests that the subject matter expertise of the authors was sound, while attention to pedagogical phrasing would benefit from earlier involvement of language specialists.

The successful validation of a 90-item bank for six specialized maritime mathematics topics addresses a practical gap in Indonesian maritime vocational education. Existing commercial adaptive learning platforms such as Khan Academy and ALEKS focus on general academic mathematics and do not address STCW-specific content. Domestic maritime polytechnics have therefore relied on ad-hoc item pools without systematic validation. The documented rubric and validation process provide a replicable model for other maritime institutions developing similar resources.

Simulation-based verification of adaptive-engine behavior adds evidence that complements expert item validation. Item-level validation shows whether individual questions meet quality standards. Simulation does something else: it checks whether the overall assessment system behaves as designed when those items are orchestrated by the adaptive algorithm. That two-pronged approach is particularly useful when large-scale learner pilots are constrained by cadet scheduling, ethical approval timelines, or limited institutional resources.

Several limitations of the current validation should be acknowledged. First, even a rigorous expert audit cannot replace empirical item response data from actual learners. Difficulty indices and discrimination indices computed from learner response data would provide stronger evidence of item quality and could reveal items that look valid to experts yet perform poorly in practice. Second, the synthetic profiles relied on static response probabilities, whereas actual cadet performance may vary across topics and within a longer training session. Finally, the content validity established here is tied to the curriculum used at Politeknik Pelayaran Surabaya. Other maritime institutions may need their own alignment checks before adopting the question bank.

These caveats point to the next stage of the study. Formal pilot tests with actual cadets are needed to gather empirical item statistics, specifically item difficulty (p-value), point-biserial correlation for discrimination, and distractor analysis. Once combined with Item Response Theory modeling, this empirical step would move the bank from a validated resource toward a calibrated instrument for high-stakes adaptive use.

CONCLUSION

This study presented a systematic validation approach for a 90-item maritime mathematics question bank developed at Politeknik Pelayaran Surabaya to support adaptive learning. It combined a structured expert pedagogical audit, organized around a four-dimension rubric, with simulation-based verification of adaptive-engine integration. The audit achieved substantial inter-rater agreement (Cohen's kappa = 0.78) and produced an overall content validity index of 0.87, with 83.3% of items accepted without modification, 14.4% requiring minor revision primarily for linguistic clarity, and 2.2% requiring major revision for technical correctness. In addition, results from 300 simulated sessions with three synthetic learner profiles showed that the 2-up/2-down staircase algorithm operated correctly and led profiles toward the expected difficulty levels. The findings contribute a validated question bank specification, a replicable validation rubric, and evidence that expert audit combined with simulation testing provides a practical validation approach suitable for specialized vocational education contexts where large-scale learner pilots are constrained. Future work will implement formal pilot testing with cadets to generate empirical item statistics and calibrate items using Item Response Theory, transforming the validated bank into a calibrated instrument for operational adaptive assessment in maritime vocational education.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131-142. https://doi.org/10.1177/0013164485451012

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer & P. A. Alexander (Eds.), Handbook of research on learning and instruction (2nd ed., pp. 522-560). Routledge.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.

Billings, D. M., & Halstead, J. A. (2019). Teaching in nursing: A guide for faculty (6th ed.). Elsevier.

Brookhart, S. M. (2010). How to assess higher-order thinking skills in your classroom. ASCD.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. Medical Education, 47(7), 726-733. https://doi.org/10.1111/medu.12202

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Kluwer-Nijhoff Publishing.

International Maritime Organization. (2017). International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW). IMO Publishing.

Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. Review of Educational Research, 86(1), 42-78. https://doi.org/10.3102/0034654315581420

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America, 49(2B), 467-477. https://doi.org/10.1121/1.1912375

Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, S. R., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments, 29(1), 142-163. https://doi.org/10.1080/10494820.2018.1558257

Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2017). Informing progress: Insights on personalized learning implementation and effects. RAND Corporation.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489-497. https://doi.org/10.1002/nur.20147

Richey, R. C., & Klein, J. D. (2007). Design and development research: Methods, strategies, and issues. Lawrence Erlbaum Associates.

Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. Psychological Bulletin, 137(3), 421-442. https://doi.org/10.1037/a0022777

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. Journal of Educational Psychology, 106(2), 331-347. https://doi.org/10.1037/a0034752

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221. https://doi.org/10.1080/00461520.2011.611369

Wahyuwono, A., Oktavia, R., & Kartikasari, D. (2024). Virtual assistant berbasis web menggunakan NLP dan LSTM untuk layanan informasi akademik. Jurnal Teknologi Pendidikan, 9(2), 201-215.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers.

Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. British Journal of Educational Technology, 50(6), 3119-3137. https://doi.org/10.1111/bjet.12700

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers in Public Health, 6, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? CBE-Life Sciences Education, 15(4), Article rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. Springer.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334. https://doi.org/10.1007/BF02310555

DeVellis, R. F., & Thorpe, C. T. (2021). Scale development: Theory and applications (5th ed.). SAGE Publications.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. Medical Education, 37(9), 830-837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). Handbook of test development. Lawrence Erlbaum Associates.

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Prentice Hall.

Gierl, M. J., & Lai, H. (2016). Automatic item generation: Theory and practice. Routledge.

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. Routledge.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 30(1), 121-204. https://doi.org/10.1007/s40593-019-00186-y

Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382-386.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412-433. https://doi.org/10.1037/met0000144

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Penfield, R. D., & Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. Measurement in Physical Education and Exercise Science, 8(4), 213-225. https://doi.org/10.1207/s15327841mpee0804_3

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. Social Work Research, 27(2), 94-104. https://doi.org/10.1093/swr/27.2.94

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. SAGE Publications.

Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three alternative item analysis indices. Educational and Psychological Measurement, 54(3), 699-706. https://doi.org/10.1177/0013164494054003011

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema, 26(1), 100-107. https://doi.org/10.7334/psicothema2013.256

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. Frontiers in Psychology, 7, Article 769. https://doi.org/10.3389/fpsyg.2016.00769

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). Elements of adaptive testing. Springer. https://doi.org/10.1007/978-0-387-85461-8

Wainer, H. (Ed.). (2000). Computerized adaptive testing: A primer (2nd ed.). Lawrence Erlbaum Associates.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Lawrence Erlbaum Associates.

Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. Education in Medicine Journal, 11(2), 49-54. https://doi.org/10.21315/eimj2019.11.2.6